APPENDIX

Details of the Systematic Review of Literature

Most cutting-edge papers on the methodology of privacy preserving record linkage have been published in either computer science or statistics journals. For our literature review, we searched the ACM Digital Library and IEEExplore, the two main databases for computer science literature. JSTOR was included to identify mathematics and statistics literature and PubMed for health related literature. Finally, ISI Web of Knowledge was selected as a good general search. ISI Web of Knowledge covers both journals and conference proceedings, which are important for capturing the computer science literature. We tested a few keyword searches to find the best keywords that were general enough to capture most of the literature, but not too broad. We then combined those keyword searches into one search as follows*: ("data integration" OR "record linkage" OR "privacy preserving record linkage" OR "private record linkage") AND "privacy."* This keyword process worked on all databases except the ACM Digital Library, where our search resulted in only 12 citations. A slight variation of (("data integration" or "record linkage") and ("privacy")) or ("privacy preserving record linkage") or ("private record linkage") resulted in 125 citations, so we used the later for the ACM Digital Library. PubMed automatically translated the keyword search to ("data integration"[All Fields] OR "record linkage"[All Fields] OR (("privacy"[MeSH Terms] OR "privacy"[All Fields]) AND preserving[All Fields] AND ("records as topic"[MeSH Terms] OR ("records"[All Fields] AND "topic"[All Fields]) OR "records as topic"[All Fields] OR "record"[All Fields]) AND ("genetic linkage"[MeSH Terms] OR ("genetic"[All Fields] AND "linkage"[All Fields]) OR "genetic linkage"[All Fields] OR "linkage"[All Fields])) OR "private record linkage"[All Fields]) AND "privacy"[All Fields].

After searching in the five databases using the keyword search, we screened the top 50 citations (sorted by relevancy in all but PubMed, where we sorted by time) for relevant articles.

We found that both PubMed and JSTOR had very few articles that covered this topic. Furthermore, the few that were identified were also found by the other databases, so we dropped citations identified by PubMed and JSTOR. We also identified an additional 22 articles from other sources. After screening the titles for duplicates, invalid citations (i.e. table of contents), and non-relevant articles, we had a total of 182 articles. For these, we reviewed the abstracts to identify methodology papers that presented specific algorithms, protocols, or architecture for carrying out privacy preserving record linkage for data integration. We ended up with 71 articles selected for full review. The main topics of papers that were excluded were privacy related database operations that were not record linkage (42), disclosure risk and reidentification (29) articles, and research perspective articles on directions for research related to record linkage and privacy (27) to support data intensive research in various areas such as epidemiology. Initially many of these perspective articles were identified as case study articles and included for full review. But we found that these articles did not include any details of protocols or experiences with record linkage in the field, so we decided to exclude them from reviewed articles. The remaining 13 papers covered a wide range of related topics such as informed consent and record linkage, evaluation of record linkage results, and record linkage methods without any specifics on privacy protection. All of these are related topics but beyond the scope of this review.

The 71 articles revealed three themes. Most methods articles, mainly published in the computer science literature, defined the computational problem as private record linkage, where the assumption is that no party has access to all data for linkage, and that the linkage function is known ahead of time. Thus, the main objective of private record linkage is to compute the known linkage function in a safe manner. However, in the case studies of actual record linkages done for research with IRB approval, the assumption of private record linkage was never met. In practice, a trusted party who was given full access to all data carried out the linkage, and the majority of the effort was spent figuring out the best linkage function for the

given databases as well as manually tuning the automatic linkages using clerical review. There was a clear gap between the cutting-edge research in methods and the needs of the biomedical informatics research community. We had screened out 27 perspective articles specifically on the need for better methods for safe record linkage to support data intensive research in many areas such as epidemiology, disability research, ambulatory care, and child maltreatment because they did not meet the inclusion criteria. However, we reviewed many of these articles and found that the use case described in the research communities did not apply to the setup for private record linkage, where two private parties are trying to link data safely. Instead, these articles refer to situations in which there is a trusted party who has full access to the data for linkage with proper approval from multiple custodians of the data sets. The challenges involved navigating the political process of getting approval from different data custodians and figuring out a trusted third party who had the trust of all parties as well as the expertise to carry out the linkage. Under HIPAA, these third parties tend to be covered entities or business associates. Researchers often work in collaboration with these entities for safe record linkage. Except for countries with dedicated linkage centers, uncertainty management of the linkage process and clerical review of the automatic linkage results is very difficult because the trusted third party, who often works for the government, has limited time to give to the researcher. The most recent articles on decoupled data linkage (SDLink) address this gap between theory and practice. Thus, we organized the review into three sections: private record linkage, case study of record linkage in practice using the trusted third party, and SDLink, which bridges the gap between the two bodies of literature. We then synthesized the literature to present a new framework: Privacy Preserving Interactive Record Linkage (PPIRL), with tractable privacy and utility properties. In the presentation of the literature, we used two separate sections, PPRL and PPIRL, for most clarity of the content. The PPRL section covered theory and practices of privacy preserving record linkage. The PPIRL section

proposed the new framework and reviewed SDLink in detail within the new framework. We wrapped up with a discussion of all literature using the new proposed PPIRL framework.

Privacy Definitions

Webster's dictionary defines privacy as *the state of being free from observation.* In the digital world, where everything is recorded and perpetual information is kept, the fundamental concept of privacy is currently in flux. Privacy in regard to personal information has been defined as the right or desire of individuals to control the release of information about themselves[1]. Controlling the release and use of digitized personal data is particularly challenging because of the ease of sharing digitized data via remote access and replication.

Privacy protection is a social construct, rather than a technical construct, that is often defined by privacy laws. Most modern information privacy laws, including The Privacy Act adopted in 1974 and HIPAA, are based on what are known as Fair Information Practices (FIPs) [2, 3]. The FIPs include five principles: *1) openness of the data system (the right to know the existence of data systems about oneself); 2) access to one's personal data; 3) integrity of one's personal data; 4) control of the use of data; and 5) security safeguards on the data*. Another commonly debated issue related to personal privacy is the right to be forgotten. In principle, EU privacy laws have some provisions for the right to be forgotten, that is, the right to request that data about oneself be permanently deleted under some circumstances. However, in a recent case, the court stated that an individual does not have a right to request deletion of accurate data, just because they don't like it. Most recently, Nissenbaum's proposal of privacy as contextual integrity, which states that privacy protection depends on the context and the expected norms of protection given a particular situation, has become widely accepted legally[4]. From a technical standpoint, these privacy standards result in policy requirements on digital data about: 1) who has access to which data; 2) for what purpose; and 3) how the data should

be maintained. The most relevant question for biomedical research is; "what is the expected norms of ethical conduct for doing research with personal data in our society." There needs to be much more public discourse, understanding, and agreement about this complex but critical question before big data can become a core resource for biomedical research.

Method	
Re-coding (Generalization, top-coding, rounding, data swapping, hashing, Bloom filters)	Mapping the value of certain attributes to reveal less information (i.e. generalize Chinese to Asian or rounding numbers).
Suppression	Not releasing the value of certain risky attributes (i.e. suppressing attributes that are rare in a table, by converting them to missing).
Obfuscation by perturbations	Masking the original data by transforming it by either adding random noise or via some form of restricted randomization. Differential privacy is most often implemented by adding random noise.
Obfuscation by Chaffing, List inflation	Adding in fake data or extra rows to the table
Sampling	Releasing only a subset of the rows
Third party linkage	Models for masked data sharing by separating the identifying data from the sensitive data, and then ensuring that no party has access to the full data at any time.
Secure multi-party computation (SMC)	A model for secure computation using multiple parties, each to compute a sub-problem with the objective that no party knows anything except its own input and the final computed results. Third party linkage discussed above is a specific example of a SMC design for record linkage.
Synthetic or simulated data	Simulating the original data to create synthetic data that is similar enough in distribution to the original data to be useful, without revealing any of the real data
Secure data centers	Restrict access to data and what analysis is allowed

Table 1. Methods used for privacy protection

Techniques for Privacy Protection

Fundamentally, privacy protection technology is about controlling the amount of information being released so that no more than the authorized information gets out. This is very difficult because often, releasing one piece of information results in other information being shared inadvertently, which can then be combined with background information to make other inferences. For example, if Alice decides to sequence her DNA and deposit it into a

public repository for research, information about her family is inadvertently made public without their consent. That is, by combining information about Alice's DNA (publicly released information) and her family information (background information), one can potentially infer sensitive health information about other members of her family that should not be released. Thus, techniques developed for privacy protection focus on revealing only the authorized information about Alice, but somehow obscuring other information about her family. Effective methods for limiting information disclosure for releasing statistical databases have been studied extensively. Table 1 summarizes the most commonly used methods.

On the technical front privacy definitions that can precisely measure the amount of information disclosed or not disclosed have been making progress. Table 2 lists the four most popular privacy definitions that are used for privacy preserving computations. Each definition has a parameter that dictates the level of information that is released, which has a direct relationship with the level of privacy protection and often an indirect relationship with the level of usefulness of the data.

Framework	Privacy Definitions (for publishing table T* derived from the private table T)
k-anonymity (1998)[5]	For every combination of quasi-identifiers (i.e. gender, date of birth, zip code, etc), called equivalence classes in the table, there are at least k records that share the same value. Thus, every record in the table is indistinguishable from at least k-1 other records with respect to the quasi-identifiers.
ℓ -diversity (2006) [6]	Each equivalence class has at least ℓ "well-represented" values for the sensitive attributes. There are different variations of what "well-representedness" means, with the simplest definition being ℓ distinct values called distinct ℓ-diversity.
t-closeness (2007) [7]	The distance between the distribution of a sensitive attribute in any equivalence class and that of the whole table is no more than a threshold t.
Differential privacy (2006) [8]	A database answers queries in such a way that the answer is insensitive to the presence or absence of an individual record in the database.

Table 2. Privacy definitions for publishing table T* derived from the private table T

REFERENCES

1. Rindfleisch, TC. Privacy, information technology, and health care. In *Commun. ACM 40, 8 (August 1997),* pages 92-100. D0I=10.1145/257874.257896.
2. Privacy Protection Study Commission. Personal Privacy in an Information Society (July 1977). http://epic.org/privacy/ppsc1977report/.
3. Kum, HC, Ahalt, S, Pathak, D. Privacy Preserving Data Integration Using Decoupled Data. Security and Privacy in Social Network, by Elovici, Y, Altshuler, Y, Cremers, A, Aharony, N, Pentland, A.(Eds), Springer 2012.
4. Nissenbaum, HF. Privacy as Contextual Integrity. *Washington Law Review* 2004;79(1):19-158.
5. Samarati, P, Sweeney, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA. 1998.
6. Machanavajjhala, A, Kifer, D, Gehrke, J, Venkitasubramaniam, M. *L*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 3 (March 2007).
7. Ninghui, Li, Tiancheng, Li, Suresh,Venkatasubramanian. *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity. In *International Conference on Data Engineering (ICDE)*, April 2007.
8. Dwork, C. Differential privacy, in: International Colloquium on Automata, Languages and Programming, 2006, pp. 1–12.